DOCUMENT RESUME

ED 281 859 TM 870 238

AUTHOR Linacre, John M.; Wright, Benjamin D.
TITLE Item Bias: Mantel-Haenszel and the Rasch Model. Memorandum No. 39.
INSTITUTION Finnish Association of Mathematics and Science Education Research.
PUB DATE Feb 87
NOTE 17p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Algorithms; Estimation (Mathematics); *Item Analysis; Measurement Techniques; *Measures (Individuals); Predictive Measurement; *Test Bias; *Test Items; *Test Theory
IDENTIFIERS *Mantel Haenszel Procedure; *Rasch Model

ABSTRACT

The Mantel-Haenszel (MH) procedure attempts to identify and quantify differential item performance (item bias). This paper summarizes the MH statistics, and identifies the parameters they estimate. An equivalent procedure based on the Rasch model is described. The theoretical properties of the two approaches are compared and shown to require the same assumptions. The MH procedure is shown to be statistically inferior to the Rasch procedure. (Author)

**********
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
**********

ED281859

# ITEM BIAS:

# MANTEL-HAENSZEL AND THE RASCH MODEL

John M. Linacre and Benjamin D. Wright

Memorandum No. 39

MESA Psychometric Laboratory

Department of Education

University of Chicago

February 1989

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

B. Wright

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Abstract:

The Mantel-Haenszel (MH) procedure attempts to identify and quantify differential item performance (item bias). This paper summarizes the MH statistics, and identifies the parameters they estimate. An equivalent procedure based on the Rasch model is described. The theoretical properties of the two approaches are compared and shown to require the same assumptions. The MH procedure is shown to be statistically inferior to the Rasch procedure.

Key words: Item bias, Mantel-Haenszel, Rasch model



TM 870 238

## Introduction

The identification and quantification of differences in item performance for contrasting groups of examinees is important if differences between groups are to be understood, and tests equivalent for groups are to be designed and maintained.

## Description of the Mantel-Haenszel procedure

Mantel & Haenszel (1959) discuss several statistical techniques for determining relative risk of disease occuring in individuals with regard to the presence or absence of other factors. One approach is to divide the samples under investigation into diseased and disease-free groups, and then match sub-categories of these groups according to the presence or absence of factors.

In their discussion of what is now referred to as the Mantel-Haenszel (MH) procedure, they explain that their intention is to address the problem of determining overall relative risk of disease as a weighted average of the relative risks in the presence or absence of various factors, with the proviso that factors which affect the risk in an extreme way are not encountered.

The MH procedure has since been proposed as an approach to detecting differences in item performance between groups differing in some other way.

After test administration, the first step of the MH procedure is to identify two examinee groups. These are the reference group, R, (chosen to provide the standard performance on the item of interest), and the focal group, F, whose differential performance, if any, is to be detected and measured. (The formulation and terminology in this paper come from Holland & Thayer (1986)). These two groups correspond to the "disease-free" and "diseased" groups.

However, there are seldom any clear external categorizing factors, so implied levels of ability are hypothesized as factors. The ability range of the groups is usually divided into three to five intervals, and these intervals are used to match samples from each group. Matching can be based on whatever information is available, which usually includes examinees' scores on the test of which the item in question is a part.

For each ability interval, of which there are now K, a 2×2 table is constructed from the responses by examinees in each sample in that interval to the target item. This table of responses made by the two sample groups in the $j^{th}$ ability interval has the form shown in figure 1.

| Sample Group | Answer made: Right (1) | Wrong (0) | Total in sample |
|---|---|---|---|
| Reference Group ($R_j$) | $A_j$ | $B_j$ | $N_{Rj}$ |
| Focal Group ($F_j$) | $C_j$ | $D_j$ | $N_{Fj}$ |
| Combined Groups: | $M_{1j}$ | $M_{0j}$ | $T_j$ |

Figure 1. Data for the $j^{th}$ matched set of members of R, the reference group, and F, the focus group.

The MH procedure is based on estimates of the probability of a member of the reference group in interval j getting the item right ($P_{Rj}$), or getting it wrong ($Q_{Rj}$), and similarly for a member of the focal group ($P_{Fj}$) and ($Q_{Fj}$).

Two statistics are derived from these estimates:

1. An estimate, $\hat{\alpha}$, of the difference in performance between the two groups across all intervals. This is an estimate of the parameter, $\alpha$, which will satisfy

$$(P_{Rj}/Q_{Rj}) = \alpha \times (P_{Fj}/Q_{Fj}) \ . \quad j=1,K \tag{1}$$

This $\alpha$ is that common odds-ratio of the two groups which is shared by each of the K 2×2 tables.

The MH equation for this performance difference estimator is

$$\hat{\alpha} = \Sigma(A_jD_j/T_j)/\Sigma(B_jC_j/T_j) \ , \quad j=1,K \tag{2}$$

in which $\hat{\alpha}$ has the range 0 to infinity with no differential performance (the null value) represented by 1.

A transformation of this statistic is proposed by Holland and Thayer to create a symmetric scale with null value of zero. This "delta scale" value is obtained by

$$\Delta = -(4/1.7) \times \ln(\alpha) \ . \tag{3}$$

According to Holland and Thayer, a proper standard error for this estimate is not yet determined though much work has been done in this area. They do include an approximation which is dependent on the number and nature of responses in each ability level and the size of the $\hat{\alpha}$ estimate.

2. An estimate of the statistical significance of the difference between the performance levels of reference and sample groups. This is a chi-square statistic with 1 degree of freedom, which, omitting correction for continuity, is

$$CHISQ = (\Sigma(A_j) - \Sigma(E(A_j)))^2 / \Sigma(Var(A_j)) , \quad j=1,K \qquad (4)$$

where

$$E(A_j) = N_{Rj}M_{1j}/T_j , \qquad (5)$$

and

$$Var(A_j) = N_{Rj}N_{Fj}M_{1j}M_{0j}/(T_jT_j(T_j-1)) . \qquad (6)$$

The MH procedure and its application to problems in the medical sphere is further discussed in Fleiss (1973) pp. 117-118 and Bishop, Fienberg, Holland (1975) pp.146-149.

What does the MH difference statistic estimate?

The practical application of the MH procedure requires an understanding of what these statistics estimate.

Consider the $\alpha$ estimated by

$$\hat{\alpha} = \Sigma(A_j D_j / T_j) / \Sigma(B_j C_j / T_j) \quad . \tag{7}$$

This estimates the parameter $\alpha$ which fulfills

$$P_{Rj}/Q_{Rj} = \alpha \times (P_{Fj}/Q_{Fj}) \quad , \quad j=1,K \tag{8}$$

where each j corresponds to an ability level

As the number, K, of ability levels is arbitrary, if this $\alpha$ is to have meaning beyond the particular matching scheme used, it must be independent of the number of levels chosen. It must also be independent of the number of pairs of examinees in each interval. In particular, it must satisfy the equation when the number of intervals is constructed to be the same as the number of pairs of examinees, with one pair of examinees in each interval.

Consequently, reformulating and taking logarithms, $\alpha$ must satisfy

$$\ln(\alpha) = \ln(P_R/Q_R) - \ln(P_F/Q_F) \tag{9}$$

for any and all pairs of examinees matched by ability

Differential item performance determined by the Rasch model

The Rasch model hypothesizes that each examinee has an ability, B, and each item has a difficulty, D. If there is differential item performance, the difficulty for the reference group $D_R$, will be different from the difficulty for the focus group $D_F$.

The items on the test other than the suspect items can be used to determine ability estimates for the sample members of both groups, (and item difficulties for all non-suspect items, if desired), according to Rasch model specifications. Procedures for performing this analysis are described in Wright & Stone (1979).

This analysis yields an ability estimate b of the ability parameter B for each examinee in each group on a common interval scale. Then, by examining performance on the suspect item, Rasch estimates are obtained for parameters which satisfy, for each member of the reference sample group,

$$B - D_R = \ln(P_R/Q_R) \ , \qquad (10)$$

and, for each member of the focus sample group,

$$B - D_F = \ln(P_F/Q_F) \ . \qquad (11)$$

Thus, for each pair of examinees who are matched on ability,

$$D_F - D_R = \ln(P_R/Q_R) - \ln(P_F/Q_F) = \ln(\alpha) \ , \qquad (12)$$

which is the formulation derived above for the ln(α) MH parameter. Since the B parameters cancel, this Rasch evaluation is independent of the distribution of abilities.

A Rasch approximation to the item bias for each interval

The item difficulty of the suspect item for the matched reference and focus groups in each interval may be estimated by the normal approximation algorithm (PROX) (Wright & Stone, 1979, Chapter 2).

For the reference group, the algorithm is

$$d_{Rj} = M_{Rj} + X_{Rj} \times \ln(B_j/A_j) \ , \qquad (13)$$

with error variance

$$s^2_{Rj} = X^2_{Rj}(A_j+B_j)/A_jB_j \ , \qquad (14)$$

where

$M_{Rj}$ is the mean ability of the reference group in interval j,

$X_{Rj}$ is a correction factor for the distribution of abilities in interval j,

and where

$$X^2_{Rj} = 1 + s^2_{BRj}/2.9 \ , \qquad (15)$$

with $s^2_{BRj}$ as the ability variance of the reference group in interval j.

For the focus group, the algorithm is similarly

$$d_{Fj} = M_{Fj} + X_{Fj} * \ln(D_j/C_j) \ , \tag{16}$$

with error variance

$$s^2_{Fj} = X^2_{Fj}(D_j+C_j)/D_jC_j \tag{17}$$

where

$M_{Fj}$ is the mean ability of the reference group in interval j,

$X_{Fj}$ is a correction factor for the distribution of abilities in interval j,

and where

$$X^2_{Fj} = 1 + s^2_{BFj}/2.9 \ , \tag{18}$$

with

$s^2_{BFj}$ as the ability variance of the focus group in interval j.

So the item bias in interval j can be estimated from

$$\frac{d_{Fj}-M_{Fj}}{X_{Rj}} - \frac{d_{Rj}-M_{Rj}}{X_{Fj}} = \ln(A_jD_j/B_jC_j) \ , \tag{19}$$

which is $\ln(\hat{\alpha}_j)$ for the $j^{th}$ interval.

Thus, if $M_{Rj}$ equals $M_{Fj}$ (the "matched" groups have equal means) and $X_{Rj}$ equals $X_{Fj}$ (the "matched" groups have equal variances),

then, when matching the $j^{th}$ interval,

$$\ln(\hat{\alpha}_j) \approx (d_{Fj} - d_{Rj}) \backslash X_{Rj} \ , \qquad (20)$$

i.e. the MH bias statistic is equivalent to the difference between the Rasch difficulty "PROX" estimates of the suspect item for reference and focus groups, adjusted by the scale coefficent $X_{Rj}$.

The standard error of this formulation of $\ln(\hat{\alpha}_j)$ is

$$S_{FRj} = \sqrt{(S^2_{Rj} + S^2_{Fj})} \ , \qquad (21)$$

which becomes, after expansion,

$$S_{FRj} = X_{Rj} \sqrt{((A_j+B_j)/A_jB_j + (D_j+C_j)/D_jC_j)} \ , \qquad (22)$$

The scale coefficent $X_{Rj}$ cancels when the test statistic for the presence of bias in interval j is formed by

$$z_j = \ln(A_jD_j/B_jC_j) / \sqrt{((A_j+B_j)/A_jB_j + (D_j+C_j)/D_jC_j)} \ . \qquad (23)$$

Thus, if the ability distributions in the $j^{th}$ interval of the reference and focus groups are approximately normal and "matched", to the extent that they have equal means and variances, then the Rasch normal approximation algorithm (PROX) can be used for estimating and testing for item bias in each interval.

Generalizing item bias across all intervals with the Rasch approach

When data fit a Rasch model, the estimates, $d_F$ and $d_R$, become independent of the ability composition of the reference group and focus group examinees.

Consequently the comparison of $D_F$ and $D_R$ does not require any matching of ability levels and consequently their estimates $d_F$ and $d_R$ can be calculated from all, or any convenient subset, of the reference and focus groups without intervals or matching.

The standard errors for $d_F$ and $d_R$ are well-defined, and calculated during the estimation procedure as $s_F$ and $s_R$. The standard error of the difference between the difficulty estimates, which measures the item bias, is

$$\text{S.E.}\ (\ln(\hat{\alpha})) = \text{S.E.}\ (d_F - d_R) = \sqrt{(s^2_F + s^2_R)}\ . \qquad (24)$$

These standard errors depend on the numbers of examinees and their ability distributions, but are independent of the size of the difference between $d_F$ and $d_R$, which determines the $\hat{\alpha}$ estimate.

Comparison of MH and Rasch approaches

The fundamental requirement of the MH procedure is that the probabilities of success for the reference and focus groups bear the same relationship across all intervals. This uniformity of relationship is required to calculate the $\hat{\alpha}$ estimate. But this calculation requires the imposition of an arbitrary segmentation and matching scheme on the two groups to be compared. Consequently, the distribution of abilities, selection of interval boundaries and the absolute sizes of reference and focus groups must affect the magnitude of the $\hat{\alpha}$ How then can this procedure estimate a parameter intended to be independent of the ability range and sample size?

The Rasch analysis builds on the same assumptions that the MH procedure implies and requires, but, by utilizing all the relevant information available from every response by the reference and focus groups, a Rasch analysis is able to provide a $\ln(\hat{\alpha})$ estimate of smaller, and better estimable, standard error, that is independent of both ability distributions.

The Rasch $\ln(\hat{\alpha})$ estimate is in "logistic units", logits, and the "delta scale", is a proportional adjustment of this logit scale.

What does the MH significance statistic estimate?

It is not enough to calculate an $\hat{\alpha}$ estimate for the performance difference $\alpha$ between any two groups. We must also evaluate the statistical significance of this difference. The MH statistic for this is

$$CHISQ = (\Sigma(A_j) - \Sigma(E(A_j)))^2 / \Sigma(Var(A_j)) , \quad j=1,K \qquad (25)$$

where

$$E(A_j) = N_{Rj}M_{1j}/T_j , \qquad (26)$$

and

$$Var(A_j) = N_{Rj}N_{Fj}M_{1j}M_{0j}/(T_jT_j(T_j-1)) . \qquad (27)$$

After algebraic manipulation, this becomes

$$CHISQ = (\Sigma((A_jD_j - B_jC_j)/T_j))^2 / \Sigma(Var(A_j)) , \qquad (28)$$

and further becomes

$$CHISQ = (\Sigma((A_j/N_{Rj})-(C_j/N_{Fj}))N_{Rj}N_{Fj}/T_j)^2/\Sigma(Var(A_j)) , \qquad (29)$$

but $A_j/N_{Rj}$ is an estimate ($p_{Rj}$) of the probability of an examinee in the reference group getting the item right, similarly $C_j/N_{Fj}$ is an estimate of $p_{Fj}$, the probability of an examinee in the focus group getting the question right.

So, the significance statistic is estimating

$$CHISQ = (\Sigma(p_{Rj} - p_{Fj})N_{Rj}N_{Fj}/T_j)^2/\Sigma(Var(A_j)) . \qquad (30)$$

Thus the MH significance statistic is obtained by averaging over different ability levels the difference between the groups of the probability of obtaining a correct response to the item.
Figure 2 shows how, for any item which has any power to differentiate between high and low ability examinees at all, and for which the difference parameter $\alpha$ is not null, $P_R - P_F$ must appear for different ability levels. Obviously no empirical mean value can represent this difference uniquely. Its size depends on the number, width and probability level of the intervals chosen. Consequently the MH CHISQ is not a stable statistic. If examinees are grouped by raw score on the test of which the item in question was a part, the arbitrary nature of the intervals may be removed, but the non-linear difference shown in figure 2 remains.

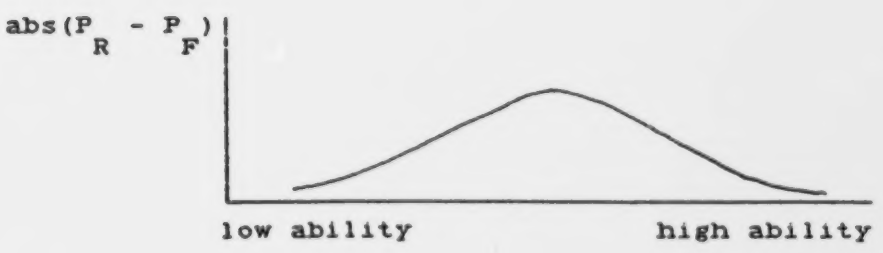


Figure 2. Difference in the absolute value of probable response to a particular item between reference and focus groups plotted against examinee ability.

## The Rasch Significance Statistic

The Rasch determination of $\ln(\hat{\alpha})$, via $d_R$ with its standard error $s_R$ and $d_F$ with its standard error $s_F$ is independent of analyst whim.

The statistics necessary to determine the statistical significance of any difference between reference and focus groups are routinely provided by Rasch analysis. The significance of an item bias can be determined by calculating the difference between the Rasch item difficulty estimates for the two groups, scaled for their standard error.
The test statistic is

$$z = (d_R - d_F) \; / \; \sqrt{(s^2_R + s^2_F)} \; , \qquad (31)$$

or, in MH terminology,

$$z = \ln(\hat{\alpha}) \; / \; \text{S.E.} \ln(\hat{\alpha}) \; . \qquad (32)$$

Note that the Rasch transformations have removed the effect of the non-linearity shown in figure 2.

## Conclusion

The Mantel-Haenszel procedure is an attempt to determine indirectly what Rasch analysis provides directly. The MH procedure involves theoretical uncertainties and depends on arbitrary decisions by the analyst who uses it.

If one is not prepared to accept the validity of the Rasch model for the item under examination, the implicit assumptions of the MH procedure will not be satisfied either. If one is prepared to accept the Rasch assumptions, however, the Rasch model yields simpler and better statistics.

References:

Bishop Y.M.M., Fienberg S.E. and Holland P.W. (1975) Discrete Multivariate Analysis: Theory and Practice Cambridge: The MIT Press

Fleiss, J.L. (1973) Statistical Methods for Rates and Proportions. New York:John Wiley & Sons

Holland, P.W. and Thayer, D.T. (1986) Differential Item Performance and the Mantel-Haenszel Procedure. Paper presented at the American Educational Research Association Annual Meeting, San Francisco, California, April 1986. (Revised April 22, 1986).

Mantel, N. and Haenszel, W. (1959) Statistical aspects of the analysis of data from retrospective studies of disease. Journal of the National Cancer Institute, 22, 719-748.

Wright, B.D. and Stone, M.H. (1979) Best Test Design. Chicago, IL: MESA Press.